Détection des experts dans un cadre incertain

Dorra Attiaoui [1], **Arnaud Martin** [2], **Boutheina Ben Yaghlane** [3]

1. *LARODEC, ISG Tunis, Université de Tunis, Tunisie*
   DRUID, IRISA, Université de Rennes 1, France
   attiaoui.dorra@gmail.com
2. *DRUID, IRISA, Université de Rennes 1, France*
   Arnaud.Martin@univ-rennes1.fr
3. *LARODEC, IHEC Carthage, Université de Carthage, Tunisie*
   boutheina.yaghlane@ihec.rnu.tn

RÉSUMÉ. *Dans cet article nous proposons une modélisation statistique des utilisateurs dans les sites communautaires de questions réponses fondée sur les théories de l'incertain. Nous utilisons une méthode de prise de décision crédibiliste fondée sur la combinaison des informations relatives à chaque type d'utilisateur. Cette approche utilise les probabilités pignistiques pour identifier chaque classe et ainsi permettre une détection des experts selon un thème spécifique.*

ABSTRACT. *In this paper we propose a statistical model for representing users in social networks and more precisely in Question Answering Communities based on uncertainty theories. The theory of fuzzy sets and the theory of belief functions allow to represent and manage uncertainty. We use their mathematical background to combine users informations and then apply an evidential decision method based on the pignistic transformation in order to classify users and detect the experts given a specific topic.*

MOTS-CLÉS : *Experts, théories de l'incertain, théorie des fonctions de croyance, classification*

KEYWORDS: *Experts,uncertainty theories, theory of belief functions, combination, classification*

DOI:10.3166/PFIA.1.1-?? © 2015 Lavoisier

1. Introduction

Dans une époque où le virtuel prime sur notre société, les gens ont acquis de nouveaux réflexes pour obtenir et consommer de l'information sur la toile. Entre sites web spécialisés, réseaux sociaux, l'utilisateur se retrouve souvent confronté à des informations qui peuvent tout aussi bien être exactes, mais parfois contradictoires, voire fausses. La récente émergence des sites communautaires de questions réponses, les a rendus très populaires auprès des internautes. Citons à titre d'exemple des sites comme Yahoo!Answers, Quora, Stackoverflow ou encore Comment ça marche. Organisés selon des thèmes et des sujets bien définis, ils permettent aux utilisateurs de poster des questions et d'y répondre. Ouverts à tous, nous nous trouvons ainsi confronté à des réponses émanant d'experts, de personnes peu formées, voire des trolls.

(Bouguessa *et al*., 2008) ont proposé que les personnes qui ont le plus de connaissances dans un domaine sont celles qui ont donné des réponses étant élues comme les meilleures dans un domaine spécifique. Cependant, (Gjergji *et al*., 2011) ont identifié trois niveaux d'incertitude dans ces sites, le premier est lié à l'extraction et l'intégration des données, le second aux sources d'information et finalement aux informations elles-mêmes. Ainsi, identifier les experts parmi des utilisateurs lambda dans un cadre incertain représente une nécessité pour que les personnes à la recherche d'information puissent obtenir des réponses fiables à leurs questions. Grâce aux représentations mathématiques offertes par les théories de l'incertain (la théorie des probabilités, des ensembles flous ou encore des fonctions de croyance), nous sommes capables de représenter et gérer les différents types d'imperfections reliées aux données.

C'est dans ce cadre que se situe notre travail, afin d'utiliser les apports proposés par ces théories pour classifier les différents types d'utilisateurs des sites communautaires et ainsi identifier les experts. Dans la suite de l'article, la section 2 présente les théories de l'incertain et plus précisément la théorie des ensembles flous et la théorie des fonctions de croyance employées dans ce travail. La section 3 s'intéresse à la détection des experts où nous présentons notre modélisation des utilisateurs, puis la comparaison de notre méthode de classification fondée sur la probabilité pignistique et celle fondée sur le K plus proche voisin crédibiliste sont présentées.

2. Les théories de l'incertain

Les théories de l'incertain sont issues de la précision des mathématiques classiques et de la subtile imprécision du monde réel. Plusieurs études ont permis d'aboutir à la théorie des ensembles flous ou des fonctions de croyance permettant de représenter l'imprécision et l'incertitude des connaissances.

La théorie des ensembles flous

La théorie des ensembles flous proposée d'abord par (Zadeh, 1965) avait pour objectif de sortir de la logique binaire en introduisant la notion d'une appartenance pondérée ou graduée. Autrement dit, permettre à un élément d'un sous-ensemble d'appartenir selon un degré plus ou moins fort à ce sous-ensemble en utilisant une fonction

appelée fonction d'appartenance. Soit un ensemble X, un sous-ensemble flou A de X est défini par : $\mu_A : X \longrightarrow [0,1]$.

La théorie des fonctions de croyance

La théorie des fonctions de croyance initialement introduite par (Dempster, 1967), formalisée ensuite dans les travaux de (Shafer, 1976) est employée dans des applications de fusion d'information et de prise de décision. A partir d'un cadre de discernement Ω ($\Omega = \{\omega_1, \ldots, \omega_n\}$) qui est l'ensemble de toutes les hypothèses, nous définissons une fonction de masse sur l'ensemble de tous les sous-ensembles possibles de Ω à qui on affecte une valeur comprise entre $[0,1]$ représentant ainsi sa masse de croyance élémentaire exprimée par $m : 2^\Omega \mapsto [0,1]$. Formellement une fonction de masse m est définie par : $\sum_{X \subseteq \Omega} m(X) = 1$.

La règle de combinaison conjonctive proposée par (Smets, 1990) est utilisée lorsque les sources d'information sont considérées comme étant fiables ou lors de l'indépendance cognitive entre les données. Elle est donnée pour tout $X \in 2^\Omega$ par :

$$m_{conj}(X) = \sum_{Y_1 \cap Y_2 = X} m_1(Y_1) m_2(Y_2) \tag{1}$$

Pour la prise de décision, la transformation pignistique permet de transformer les fonctions de masse en mesures de probabilité. Ces fonctions permettent ainsi une prise de décision sur seulement des singletons. La probabilité pignistique définie par :

$$BetP(X) = \sum_{Y \in 2^\Omega, Y \neq \emptyset} \frac{|X \cap Y|}{|Y|} * \frac{m(Y)}{1 - m(\emptyset)}, \forall X \in 2^\Omega, X \neq \emptyset \tag{2}$$

3. Modélisation

Dans cette section nous décrivons le modèle proposé fonder sur l'utilisation des théories de l'incertain présentée dans la section 2. La modélisation des utilisateurs est fondée sur les fonctions d'appartenance permettant de représenter l'imprécision des informations. De ce fait, nous serons ainsi en présence d'un modèle imprécis offrant une vision imparfaite de la réalité. Ensuite, le passage aux probabilités, offre la modélisation de l'incertitude sur le modèle. Ces théories aussi riches soient elles sont des cas particuliers de la théorie des fonctions de croyance. Ainsi, cette dernière offre un cadre adéquat permettant de modéliser aussi bien, l'incertitude, l'imprécision que l'ignorance. C'est donc naturellement que nous avons choisi de construire notre travail sur ce formalisme qui offre un outil performant pour la fusion d'information. Nous considérons quatre types d'utilisateurs sur les sites de questions-réponses :

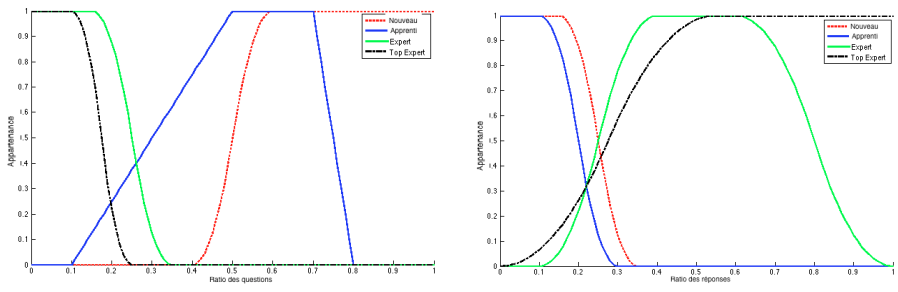
1. *Nouveau* : qui commence à apprendre les notions de base
2. *Apprenti* : qui a connaissances des notions de bases
3. *Expert* : donnant des réponses satisfaisantes et de bonne qualité
4. *Top Expert* : qui a de grandes connaissances et des réponses de bonne qualité

Nous reprenons ici l'hypothèse proposée par (Zhang *et al.*, 2007) considérant le nombre de questions et de réponses comme étant des indicateurs de l'expertise d'une personne. Ils se fondent sur le fait que plus une personne pose des questions, plus elle manque de connaissances dans un domaine. Ainsi, dans ce cas, un utilisateur est considéré comme étant un *Nouveau* ou un *Apprenti*. Par ailleurs, plus un utilisateur fournit de réponses, plus il a d'expertise, et donc qu'il soit un *Expert* ou un *Top Expert*. Par conséquent, la modélisation de chaque utilisateur est directement reliée à deux variables relatives aux nombres de questions et de réponses. Ces variables sont mesurées à partir de deux rapports :

1. Rapport des questions : *Nombre de questions posées par utilisateur par thème*/ *Nombre total des questions posées par thème*
2. Rapport des réponses : *Nombre de réponses données par utilisateur par thème*/ *Nombre total des réponses données par thème*

Ici, nous proposons une modélisation probabiliste fondée sur les fonctions d'appartenance pour décrire le comportement des utilisateurs comme présenté dans la figure 1 selon les rapports des questions à droite et des réponses à gauche. Les fonctions d'appartenance permettent une modélisation précise du degré d'appartenance d'un utilisateur à une des classes.

A partir de ces fonctions d'appartenance, nous calculons les probabilités de chaque distribution décrivant le comportement des utilisateurs. Par la suite, nous construisons les fonctions de masse à partir du principe du moindre engagement. Ce dernier nous encourage à choisir la masse qui entraîne le moins de conséquences afin de ne pas présumer d'information que nous ne possédons pas. C'est une approche naturelle qui nous engage le moins possible vis à vis du choix des croyances. Les deux fonctions de masse ainsi obtenues une pour les questions l'autre pour les réponses sont combinées par la combinaison conjonctive de l'équation (1) de façon à renforcer les informations concordantes. Nous supposons ici l'indépendance cognitive entre le nombre de questions posées et le nombre de réponses données. La décision est ensuite prise par la probabilité pignistique offrant un compromis entre les différentes règles existantes.



(a) Représentation des Questions (b) Représentation des Réponses

FIGURE 1 – Fonctions d'appartenance

Nous avons modélisé les réponses et les questions en fonction du type d'utilisateur par les fonctions d'appartenance données dans la figure 1. Dans la figure 1a relative aux fonctions d'appartenance selon le rapport des questions, les deux premières courbes relatives aux *Experts* et aux *Top Experts*, nous avons fixé le degré d'appartenance pour un faible nombre de questions a une valeur maximale de $\mu = 1$, puis elle régressent progressivement jusqu'à avoir une valeur nulle. Ceci modélise qu'à partir d'un certain rapport entre les questions posées et le nombre total par thème, aucun utilisateur présumé appartenir à ces classes ne pose de question. Nous modélisons le comportement inverse pour les *Apprentis* et les *Nouveaux*. À partir d'un rapport de 0.7 un *Apprenti* ne pose plus de questions contrairement aux utilisateurs de la classe *Nouveau* qui pour un rapport de 1 ont une appartenance de 1. Nous conservons le même raisonnement pour les réponses mais avec une représentation différente comme décrit dans la figure 1b où l'appartenance d'un *Top Expert* a une valeur maximale de 1 et inversement pour les *Nouveaux* et les *Apprentis*.

4. Expérimentations

Nous avons généré : 999 *Nouveau*, 402 *Apprenti*, 101 *Expert*, 52 *Top Expert*, de façon à simuler un petit nombre d'experts et de top experts reflétant la réalité. Contrairement aux *Nouveaux* et aux *Apprentis* dont le nombre est relativement élevé car ils sont très nombreux sur les sites communautaires.

Nous procédons dans un premier temps à la classification des utilisateurs en employant la probabilité pignistique de l'équation (2). La matrice de confusion obtenue est présentée dans le tableau 1. Nous avons obtenus les probabilités de classification correcte avec une valeur de 0.8729 pour les *Nouveaux*, 0.5970 pour les *Apprentis*, 0.8020 pour les *Experts* et enfin 0.6154 pour les *Top Experts*.

Tableau 1 – Matrice de confusion avec BetP

	Nouveau	*Apprenti*	*Expert*	*Top Expert*
Nouveau	871	127	0	0
Apprenti	159	240	3	0
Expert	0	0	81	20
Top Expert	0	0	20	32

Afin d'évaluer notre méthode de classification, nous la confrontons au classificateur crédibiliste KNN de (Denoeux, 1995). La matrice de confusion est donnée dans le tableau 2. Nous avons obtenu les probabilités de classification correcte, avec une 0.8118 valeur pour les *Nouveaux*, 0.6517 pour les *Apprentis*, 0.7624 pour les *Experts* et enfin 0.4808 pour les *Top Experts*.

En comparant les résultats des tableaux 1 et 2 respectivement fondés sur la BetP et le E KNN, nous remarquons que notre méthode permet un meilleur taux de détection des utilisateurs de types *Experts* et *Top Experts* (0.8020 et 0.6154 avec le BetP contre

Tableau 2 – Matrice de confusion avec E KNN

	Nouveau	*Apprenti*	*Expert*	*Top Expert*
Nouveau	811	188	0	0
Apprenti	136	262	2	2
Expert	0	0	77	24
Top Expert	0	0	27	25

0.7624 et 0.4808 avec E KNN). Par ailleurs, nous remarquons que pour la détection des *Apprentis*, la deuxième méthode est légèrement meilleure avec une classification de 0.6517 et seulement 0.5970 pour la BetP. Ainsi, la classification fondée sur la BetP présente de meilleurs résultats pour la détection des *Experts* et des *Tops Experts* par rapport au K NN crédibiliste.

5. Conclusion

Dans cet article, nous avons proposé une représentation statistique fondée sur les théories de l'incertain des utilisateurs dans les réseaux communautaires. Nous avons combiné les informations relatives aux nombres de questions et réponses données pour chacun d'entre eux afin de permettre une prise de décision et une classification de chaque type d'utilisateur. Cette méthode fondée sur les probabilités pignistiques nous a permis d'avoir un meilleur taux de détection des experts que le E KNN. Comme perspectives à ce travail, nous allons non seulement explorer la notion des votes pour choisir les meilleures réponses mais aussi nous concentrer sur l'application de ce modèle sur des données réelles pour valider l'approche proposée.

Bibliographie

Bouguessa M., Dumoulin B., Wang S. (2008). Identifying authoritative actors in question-answering forums: the case of yahoo! answers. In, p. 866-874. ACM.

Dempster A. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, vol. 38, p. 325-339.

Denoeux T. (1995). A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, p. 804-813.

Gjergji K., Jurgen V G., D. S., Thore G. (2011). Cobayes: bayesian knowledge corroboration with assessors of unknown areas of expertise. In, vol. 2, p. 465-474.

Shafer G. (1976). *A mathematical theory of evidence*. Princeton University Press.

Smets P. (1990). The Combination of Evidence in the Transferable Belief Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, n° 5, p. 447–458.

Zadeh L. A. (1965). Fuzzy sets. *Information and Control*, vol. 40, n° 8, p. 338–353.

Zhang J., Ackerman M. S., Adamic L. (2007). Expertise networks in online communities: structure and algorithms. In, p. 221–230. ACM Press.